Evaluating and Evolving Metadata in Multiple Dialects

Kozimor, John
jkozimor@hdfgroup.org
Ted Habermann [1] | Sean Gordon [1] | Lindsay Powers[2]
IN23C-1781

QUICK EVALUATION REPORT

Quick Evaluation reports provide an empirical analysis of fields (XML) organizations actually use in metadata collections. They are ideal for comparing metadata across collections.

The ESDIS Unified Metadata Model recommends including metadata about five kinds of people and organizations. The first two reports compare how metadata contact and distributor contact are implemented. The third report compares usage of email addresses across all five contact types. Numeric values show the average number of occurences of each element / record. Pink cells indicate missing elements. See https://wiki.earthdata.nasa.gov/display/NASAISO/NASA+ISO+for+EOSDIS for other examples.

Metadata Contact

Elements

Collections

Concept	Element Path	ASF	CDDIS	GES_DISC	GHRC	GSFCS4PA	LAADS	LANCEAMSR2	LANCEMODIS	LARC	LARC_ASDC	LPDAAC_ECS	NSIDC_ECS	OB_DAAC	OMINRT	ORNL_DAAC	PODAAC	USGS_EROS
gmd:organisationName	gmd:contact/gmd:organisationName	1.00	1.00		1.00	1.00		1.00	.34	.45	1.00	1.00	.93	1.00	1.00	1.00	1.00	1.00
gmd:CI_RoleCode	gmd:contact/gmd:role/gmd:CI_RoleCode	1.00	1.00		1.00	2.01		1.00	.34	.45	1.00	1.00	.93	1.00	1.00	1.00	1.00	1.00
gmd:individualName	gmd:contact/gmd:individualName					1.01												
gmd:deliveryPoint	gmd:contact/gmd:contactInfo/gmd:address/gmd:deliveryPoint					1.00												
gmd:city	gmd:contact/gmd:contactInfo/gmd:address/gmd:city					1.00												
gmd:administrativeArea	gmd:contact/gmd:contactInfo/gmd:address/gmd:administrativeArea					1.00												
gmd:postalCode	gmd:contact/gmd:contactInfo/gmd:address/gmd:postalCode					1.00												
gmd:country	gmd:contact/gmd:contactInfo/gmd:address/gmd:country					1.00												
gmd:voice	gmd:contact/gmd:contactInfo/gmd:phone/gmd:voice					1.38												

Differences between collections reflect different approaches to providing contact information.

Distributor Contact

Elements

Collections

Concept	Element Path	ASF	CDDIS	GES_DISC	GHRC	GSFCS4PA	LAADS	LANCEAMSR2	LANCEMODIS	LARC	LARC_ASDC	LPDAAC_ECS	NSIDC_ECS	OB_DAAC	OMINRT	ORNL_DAAC	PODAAC	USGS_EROS
gmd:CI_RoleCode	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:role/gmd:CI_RoleCode	1.00	1.00	1.00	1.00	1.11	1.00	1.00	1.00	1.36	2.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
gmd:voice	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:phone/gmd:voice	1.00	2.00			2.21			1.03	.79	1.00	1.00	.93			2.00	1.00	4.00
gmd:organisationName	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:organisationName		1.00	1.00					.66	1.00	2.00	1.00	.99	1.00	1.00		1.00	1.00
gmd:deliveryPoint	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:address/gmd:deliveryPoint		1.00			1.11			.34	.79	1.00	1.00	.93			1.00		1.00
gmd:city	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:address/gmd:city		1.00			1.11			.34	.79	1.00	1.00	.93			1.00		1.00
gmd:administrativeArea	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:address/gmd:administrativeArea		1.00			1.11			.34	.79	1.00	1.00	.93			1.00		1.00
gmd:postalCode	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:address/gmd:postalCode		1.00			1.11			.34	.79	1.00	1.00	.93			1.00		1.00
gmd:country	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:address/gmd:country		1.00			1.11			.34	.79	1.00	1.00	.93			1.00		1.00
gmd:individualName	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:individualName		1.00			1.11				.35			.01			1.00	1.00	
gmd:contactInstructions	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:contactInstructions								.34	.79		1.00	.92					
gmd:hoursOfService	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:hoursOfService									.78		1.00	.93					
gmd:positionName	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:positionName									.35			.01				1.00	
gmd:facsimile	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:phone/gmd:facsimile									.78		1.00	.93					
gmd:electronicMailAddress	gmd:distributionInfo/gmd:distributor/gmd:distributorContact/gmd:contactInfo/gmd:address/gmd:electronicMailAddress						.21		.05									

Distributor contacts are more heterogeneous across collections and many include mailing information.

Email

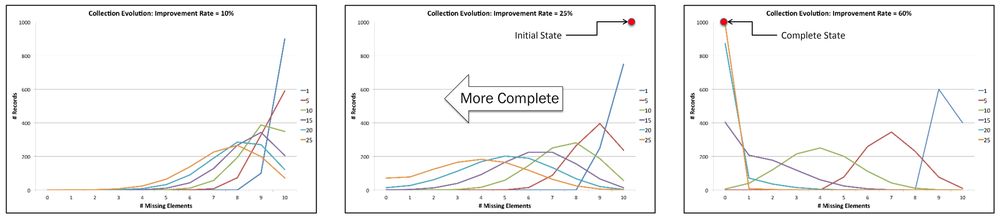
Elements

Collections

Concept	ASF	CDDIS	GHRC	GSFCS4PA	LAADS	LANCEMODIS	LARC	LARC_ASDC	LPDAAC_ECS	NSIDCV0	NSIDC_ECS	OMINRT	PODAAC	SEDAC	USGS_EROS
Metadata Contact				.91											
Distribution Contact	1.00	1.00	1.00	1.14	1.00	1.00	1.32	1.00	1.00		1.00		.99	1.00	
Resource Author / Originator				.95	.98	1.00	.48		.94	.03	1.68				
Point of Contact				.57											
Processor															

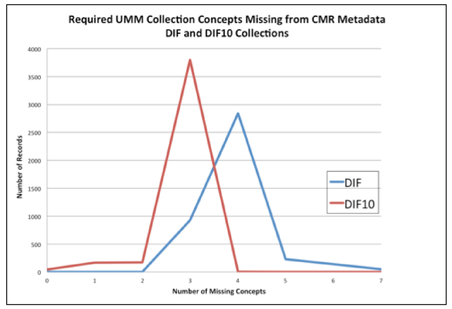
E-mail is the most common contact mechanism in use today but this information is missing from many collections.

COLLECTION EVOLUTION



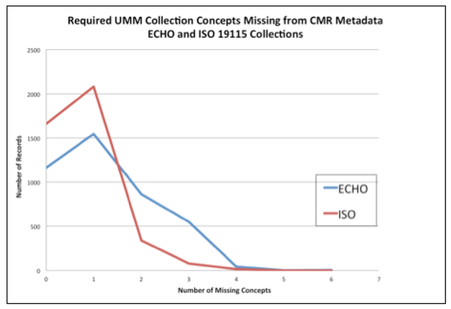
We are interested in understanding how metadata collections evolve towards completeness over time and characterizing that evolution graphically. These plots show evolution of a collection of 1000 metadata records that are each missing 10 elements initially (red circle in middle frame). We assume that some % of the missing elements are supplied during each time step. These frames show increasing improvement rates from left to right (10%, 25% and 60%) and distributions after different time steps (1-25). The distributions move towards the left as the number of missing elements decreases and eventually will end up in the upper left corner of the plot (no missing elements in any records). The 25th time step in the third frame shows this final state. Two real-world examples below show collection improvements driven by two different kinds of change.

DIF Dialect Evolution



Collection completeness with respect to a given recommendation can be improved by extending the dialect to include concepts that are missing in the original dialect. This approach was used in updating the Directory Interchange Format (DIF) dialect to include more concepts from the Unified Metadata Model recommendations.

ISO to ECHO Translation Evolution



Collection completeness with respect to a given recommendation can be improved by translating the metadata to a dialect that is more complete. This approach was used in translating metadata records from the EOSDIS Clearinghouse (ECHO) dialect to the ISO 19115 dialect which is more complete with respect to the Unified Metadata Model recommendation.

RECOMMENDATION ANALYSIS DASHBOARD

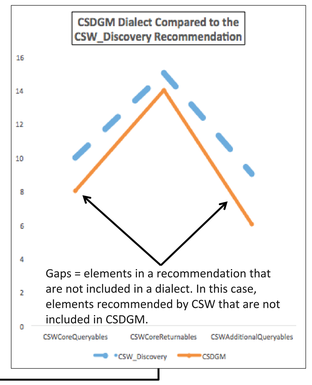
The Recommendation Analysis Dashboard is an exploratory metadata evaluation tool. It enables metadata for a single dialect to be easily evaluated using multiple recommendation, such as OGC Catalogue Services for Web (CSW) or Data Citation (DataCite).

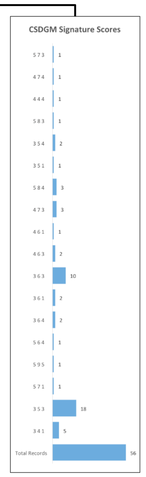
DIALECT SUITABILIY

Many recommendations are closely associated with specific dialects so gaps may emerge when used with a different dialect. Identifying gaps between dialect capabilities and recommendation requirements is an important first step in the evaluation.

SIGNATURE SCORES

Records that are missing the same number of elements (typically the same fields) are called “signature groups”. The signatures include a digit that gives the number of missing elements for each level of a recommendation (3 in this case). A complete record has a score of 0 0 0.





Collection Concept	Occurance
Concept	CSDGM	Recommendation Profile
Keyword	86%	CSWCoreQueryables
Resource Title	100%	CSWCoreQueryables
Abstract	100%	CSWCoreQueryables
Resource Format	0%	CSWCoreQueryables
Metadata Identifier	-100%	CSWCoreQueryables
Metadata Modified Date	100%	CSWCoreQueryables
Resource Type	0%	CSWCoreQueryables
Bounding Box	71%	CSWCoreQueryables
Coordinate Reference System (CRS)	0%	CSWCoreQueryables
Association	-100%	CSWCoreQueryables
Resource Title	100%	CSWCoreReturnables
Author / Originator	98%	CSWCoreReturnables
Keyword	86%	CSWCoreReturnables
Abstract	100%	CSWCoreReturnables
Publisher	18%	CSWCoreReturnables
Contributor Name	54%	CSWCoreReturnables
Metadata Modified Date	100%	CSWCoreReturnables
Resource Type	0%	CSWCoreReturnables
Resource Format	0%	CSWCoreReturnables
Metadata Identifier	-100%	CSWCoreReturnables
Source Citation	2%	CSWCoreReturnables
Metadata Language	100%	CSWCoreReturnables
Related Resource Citation	0%	CSWCoreReturnables
Bounding Box	71%	CSWCoreReturnables
Rights	100%	CSWCoreReturnables
Resource Revision Date	-100%	CSWAdditionalQueryables
Resource Creation/Revision Date	18%	CSWAdditionalQueryables
Publication Date	18%	CSWAdditionalQueryables
Organization Name	98%	CSWAdditionalQueryables
Security Constraints	0%	CSWAdditionalQueryables
Metadata Language	100%	CSWAdditionalQueryables
Resource Identifier	-100%	CSWAdditionalQueryables
Parent Identifier	-100%	CSWAdditionalQueryables
Keyword Type	80%	CSWAdditionalQueryables

-100% = Concept Not in Dialect
0% = Concept Not in Collection
100% = Concept in All Records
54% = Concept in Some Records

Concept	Guidance	Links
Incomplete use in the Collection	Unused in Collection	Missing from Dialect
Keyword		
	Resource Format	Metadata Identifier
Bounding Box	Resource Type	Association
Author / Originator	Coordinate Reference System (CRS)	Resource Revision Date
Publisher	Related Resource Citation	Resource Identifier
Contributor Name	Security Constraints	Parent Identifier
Source Citation		
Resource Creation/Revision Date		
Publication Date		
Organization Name		
Keyword Type		

RESULTS SUMMARY

List all of the elements in the recommendation and show how often they occur in the metadata collection. Identify fields that are missing from dialect, missing from collection, complete, or partial.

CONCEPT GUIDANCE

Links to online guidance in the ESIP wiki for incomplete, unused or missing concepts. See http://wiki.esipfed.org/index.php/Category:Documentation_Connections

See https://www.slideshare.net/secret/D6ZbE1t55FMzyK for more information

This material is based upon work supported by the National Aeronautics and Space Administration under Raytheon Contract Number **NNG15HZ39C**
NASA Official: Mcinerney, Mark A. (GSFC-4230) | mark.mcinerney@nasa.gov

[1]The HDF Group | [2]U.S.G.S.